Article Title: Old Russian Texts and Computer-Controlled Lexicography

Article Author: Dietze, Joachim

Issue Date: December 1987

Publisher: William R. Veder, Slavisch Seminarium, Universiteit van Amsterdam, Postbus 19188, 1000 GD Amsterdam (Holland)

Citation: *Polata Knigopisnaia: an Information Bulletin Devoted to the Study of Early Slavic Books, Texts and Literatures* 17-18 (December 1987): 125-129.

OLD RUSSIAN TEXTS AND COMPUTER-CONTROLLED LEXICOGRAPHY

JOACHIM DIETZE

Director, Universitäts- und Landesbibliothek Sachsen-Anhalt
August-Bebel-Str. 13, 4010 HALLE (Saale)
DDR

Medieval manuscripts are absolutely important sources for historical linguistic research, i. e. they are the essential basis of investigations on the history of a literary language, for literary languages are characterized by the following features, which scientists can determine in literary manuscripts: language is structured to a certain degree and therefore tends to be standardized. Besides it is important that literary languages have a wide spectrum of functions. This multivalency becomes clearly visible by the character of medieval manuscripts being monuments of the literary language. In close connection with it languages are stilistically differentiated, so that one can define different functional styles. Investigating manuscripts written in a medieval literary language requires a great deal of energy, because you ought to consider all the single linguistic phenomena, in order to compare them both diachronically and synchronically. Only then you'll really determine linguistic structures and can follow their development. This work can be made easier with the help of computers, which can be a worthful aid of diachronic language research. First of all they are able to register lexemes needed for graphematic, phonological and morphological investigations. However, it must be pointed out that processing the single linguistic levels makes different demands to data processing. Using computers gives the following advantages to lexicographers:[1)]

1. Many operations can be performed quickly and without any human power.
2. There is no loss of information on linguistic material.
3. Computers allow to use several possibilities of sorting.
4. Numerical values can be counted and calculated automatically.
5. The computer output, i. e. the data medium can directly be used for publications.

Using computers as automatic lexicographers it is an obvious thing to investigate the frequency of lexemes being an important statistical index. By this reason it is recommended to make frequency dictionaries. Choosing the texts which frequency dictionaries will be based on, you must care about their quality and quantity. Quality means the properties of a text, i. e. the type of language, the style, the theme and the date of its origin. Comparative philological investigations can be conducted only on texts of equal quality. Quantity concerns the text size needed in order to obtain statistically representative results. If you investigate a literary monument as a whole, then there are no problems with quantity. However, choosing single texts, you must take care that the sample is big enough. The relative error or the standard deviation can be an indicator for it, where the last one shows the mean variation of the factor investigated.

Diachronic language research needs monolingual frequency dictionaries with a certain structure. Making them we normally start from one single literary monument, i. e. medieval manuscript, in order to carry out graphematic, phonological, morphological and word-formation investigations. Such a frequency dictionary should absolutely contain the single word forms of the literary monument and their lemmata. The computer registers all the word forms of the text and expresses their occurrencies by frequency-statistical indices without any difficulties. On the other hand,

determining lemmata of historical texts is rather difficult, because Old Russian doesn't have any standardized orthography. Unfortunately, the possibilities of automatical lemmatization of modern texts cannot be applied to diachronic research. That's why we must conduct all the investigations above all by hand. Word forms and lemmata can be grouped by the computer in alphabetic order, in retrogressive alphabetic order (a tergo) and after diminishing frequency. For linguostatistic investigations it is recommended to add to the lexemes both their absolute and their relative frequency rate (related to the number of lexemes in the whole corpus).

Reading-in handwritten texts into the computer, you must observe certain principles of edition, otherwise you cannot obtain the lexicographic effects wanted. In detail you have to fulfill the following obligatory demands:

1. In order to conduct graphic and phonological investigations you must maintain the graphic peculiarities of the texts. That means, the graphemes must not be modificated, but allographs being phonetically irrelevant orthographical variants must be substituted by adequate graphemes.
2. The words must be distinguished one from another on uniform principles. Special attention must be paid to the problems of proclisis and enclisis and of the formation of compound words. In such a case you have to add enclitic words, if you notice a semantic change in comparison with the original simplexes, then you have a compound word.
3. Paleographic abbreviations must be resolved (per contractionem or per litteras superpositas). That is valid, too, for word forms shortened by proclisis and enclisis.
4. Least of all lexemes of low frequency must be characterized by their

place in the text, i. e. by their position in the literary monument (at least by the page number). Attention must be paid to the fact that if the beginning of a page lays in the middle of a word, you must artificially transfer it behind the end of the word, in order to avoid to divide words. Marking the place of a lexeme, you'll ask himself how to segment the whole text. It may be most easy merely to keep the original structure of the manuscript, i. e. to divide the text into pages and lines.

5. Evident spelling mistakes must be corrected.

The following principles of editing texts for reading them into a computer are of facultative nature:

1. Proper names are an important subject of linguistic research on the history of languages, because proper names are spread and used under special laws. That's why you ought to emphasize them in the text by capitalization. This formal marking enables us to register them separately as lexemes.
2. The word is of great meaning not only for syntactical research, but also for invesigation in the field of morphology and word formation. By this reason it is recommanded to add at least to the lemmata any word class information. In Old Russian texts you can partially automate this operation with the help of derivation and flexion suffixes.

The monolingual frequency dictionaries of Old Russian texts (First and Fourth Novgorod Chronicle)[2] edited by us according with the above-mentioned principles allow to conduct comparative diachronic investigations of the Russian literary language on a linguostatistical basis.

Annotations

1) Dietze, J.: Das einsprachige Frequenzwörterbuch für die diachronische Sprachbetrachtung - Aufbau und rechnergestützte Herstellung. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung. 35 (1982), p. 299.

2) Dietze, J.: Frequenzwörterbuch zur Synodalhandschrift der Ersten Novgoroder Chronik. Halle (Saale) 1977. (Martin-Luther-Universität Halle-Wittenberg. Wissenschaftliche Beiträge. 1977, 13 (F 11).)
Dietze, J.: Frequenzwörterbuch zur jüngeren Redaktion der Ersten Novgoroder Chronik. München 1984. (Sagners slavistische Sammlung.5.)
Dietze, J.. Frequenzwörterbuch zur Vierten Novgoroder Chronik. Halle (Saale) 1984. (Martin-Luther-Universität Halle-Wittenberg. Wissenschaftliche Beiträge. 1984, 17 (F 49).)